The effectiveness of EBSE in modifying professional practice – An experimental protocol and Study

Graham Fletcher and Marie Cahillane
Cranfield University, Shrivenham, UK.
g.p.fletcher@cranfield.ac.uk
m.cahillane@cranfield.ac.uk

Abstract

This paper reports on an experiment to investigate if evidence based systematic reviews have the power to modify people's opinion. A blind, primary study was conducted indicating that there are statistically significant differences between people's opinions when compared with coving the same material in an unstructured manner.

1 Introduction

Evidence based software engineering (EBSE) is the extension of evidence based scientific principles into the field of software engineering. Evidence based science comes primarily from medical research where it was discovered that following the recommendations of systematic reviews provides better results than those achieved by following expert advice alone (Antman 1992). The process of systematic review is central to EBSE. In one of the first published papers to propose the concept Kitchenham et. al. state

> "...some aspects of EBSE are essentially low risk and should be adopted as soon as possible, such as the development and adoption of guidelines for systematic reviews." (Kitchenham 2004)

Since 2004 a significant amount of research effort has gone into the development of systematic reviews in a wide range of software engineering topics (Kitchenham 2009). However, the point of a systematic review is to affect practice. There is little point in completing work that has little or no effect on the software engineering practitioners that use and implement research output. In their paper on the teaching of EBSE Jrgensen et. al. state

> "We cannot claim that we have demonstrated that teaching EBSE has a significant positive effect on real world software development work....There are no scientific studies on the effects of EBSE. The empirical evidence regarding its effectiveness is constituted only by our own practice...." (Jorgensen 2005)

They argue that by analogy to medical sciences, teaching of EBSE techniques should have a small but positive impact on the quality of software engineering practice. This may or may not be true; however, it is not an argument for producing EBSE reviews as inputs into professional decision making. There is a significant difference between a professional engineer understanding the concepts of EBSE and performing systematic reviews when faced with a complex decision and the same engineer accepting the findings of externally produced reviews. We argue that medical science faces the same issues. In the UK this has been addressed by the formation of the National Institute for Clinical Excellence. This statutory body is embarking on a process of mandating practice according to fixed guidelines and is forcing the adoption of systematic review findings.

Without a statutory body to enforce the adoption of findings it is necessary to understand the power of EBSE to affect practice. In particular, we wish to know if systematic reviews have the ability to affect

practice. If not, should we concentrate our efforts in teaching the general techniques to engineers and allow them to perform their own analysis.

There are many complex issues to address in this area. This paper describes a protocol we have adopted to measure the effectiveness of a systematic literature review in affecting the opinion, confidence level and plans of decision makers within the UK Ministry of Defence (UK MoD).
The hypotheses for our research questions are as follows:

1. Exposure to a systematic review has the power to change the opinion of knowledgeable participants.
2. Systematic reviews produce a narrower range of opinion than those formed by participants only exposed to primary material.
3. Exposure to a professionally completed systematic review increases the perceived confidence participants have in their opinions.

2 Experimental Protocol

We use a modern and much discussed proposal in software engineering as our research tool, pair programming. The concept of pair programming is that two people will share one terminal and work together to increase quality, reduce faults and depending on the research project even reduce elapsed time and total cost. We selected this topic as it is controversial, the subject of many primary studies and has been used as a topic for systematic literature review (Hannay 2009).

Our experiment is a primary study that investigates the affect of the independent variable of exposure / non-exposure to a systematic review" on the following dependant variables using a within-subjects, before and after group design:

- Efficacy: The participant's opinion of the utility of pair programming.
- Confidence: The confidence level that the participants have in their understanding of pair programming.

This design allows the comparison of the data from the pre-exposure measure of the dependent variables with the post-exposure measure of the dependent variables. Within group differences were controlled, allowing any change in the dependent variables to be identified.

2.1 Procedure

1. Independently and with no knowledge of this study by either the participants or the trainer, the participants were introduced to pair programming. All the background material on pair programming used by Hannay et. al. (Hannay 2009) was presented and formed the basis of a discussion on the utility of pair programming. This training is a normal part of the staff development provided by the UK MoD and was presented by a leading exponent of pair programming.

2. Fourteen days after the initial training the participants were asked to complete a questionnaire which provided a measure of their opinions on pair programming. The participants were also asked to rate how confident they were in the responses they had given in the questionnaire.

3. The participants were then introduced to the concepts of EBSE and systematic review. As part of this introduction several concepts were revisited, including pair programming. The participants were asked to review the way they had formed their initial opinions by rating their level of confidence in their responses to items in the questionnaire used earlier.

4. During the last stage, the participants were exposed to a professionally completed systematic literature review (Hannay 2009) and it's findings were analysed and discussed. The questionnaire completed during the initial participant training phase was re-completed, and again, as at stages two and three, the participants were asked to rate their level of confidence in their responses to the questions. Participants were asked to voluntarily submit their results to the study after the completion of the third stage of the study.

The following questions form the questionnaire completed by participants:

1. Is pair programming a worthwhile methodology?

2. Does pair programming effect software quality?
3. Does pair programming effect speed?
4. Does pair programming effect costs?
5. In what situation is pair programming particularly effective?
6. In what situation is pair programming particularly in-effective?

Data was initially collected as handwritten questionnaires and collected so that the participants were not aware of the study during its operation, but they become fully informed before agreeing to submit their results. Answers submitted by participants were coded so that it was not possible to identify individuals or to identify which subjects declined to include their results. A nine-point Likert scale was used for answers to all questions except 5 and 6 which have free form answers. An Independent analysis committee, blind to the hypotheses, performed a thematic analysis of the free form answers. The categories were then mapped to nine-point Likert scale answers by subject experts who were asked to score each category as indicative of an answer to the question "Is pair programming a worthwhile methodology? Most of the themes were scored remarkably consistently by the expert panel as indicated by very low standard deviations. Where there was no consistent opinion of the meaning of a category, indicated by a standard deviation of above 1.5, the answers falling into this category were removed from the study.

The following dependant variables were derived from the data gathered.

1. Efficacy: The expressed opinion on the efficacy of pair programming as a technology. Calculated as the average likert score from questions 1 to 6. In the open questions (5 and 6) the Likert scores were derived by the thematic analysis and independent analysis.
2. Confidence: The expressed confidence, on the same nine-point Likert scale, of the validity of the subject's answers.
3. Intent: The expressed intent to deploy pair programming on the projects or programme for which the subject is responsible. Responses are on the nine-point Likert scale. Where no response was given then the subject was omitted from the study for the elements that require this variable.

2.2 Data Analysis Procedures

In this section we detail our approach to analysing our data in preparation for and demonstrate how this data is used to answer our research hypotheses. The overall analysis procedure was to use statistical bootstrapping techniques to calculate the probability that the data sets gathered before and after exposure to systematic reviews were from different populations. This class of statistical re-sampling techniques provides a robust method of evaluating the statistical significance of any changes to the variables in the before and after populations.

2.2.1 Hypothesis 1: Exposure to a systematic review will change the opinion of knowledgeable participants.

In this question we are looking at the population of subjects. We wish to know if exposure to a systematic review has the power to change the opinion of the group by a measurable, significant amount. In this question we are dealing with perceived efficacy and confidence. I.e. we see a change of efficacy and confidence as equally important.

Expressed and true efficacy values are assumed to differ according to a standard distribution, with the expressed value at the centre of the distribution and the confidence related to the deviation. The distributions are calculated using the probability density function for a standard distribution, normalised such that the probability falling in the Likert range (1-9) is 1. Table 1 and Figure 1 set out the standard deviations used for each expressed certainty.

The probability distributions of a population of participants can be summed to produce a graph representative of the whole population. Each of the individual graphs has an area under the curve of 1, summing *n* graphs gives a "probability distribution curve" representative of *n* answers with an area under the graph of n. Given two "population curves" and integrating the absolute distances between

them over the valid Likert responses gives a distance between the populations as required by the statistical resampling.

This technique has several major advantages. Two participants giving expressed efficacies of 5 and 6 will be much closer than responses of 1 and 9. A straight histogram would show them as equally distant. It also quantifies the difference between efficacies of 5 expressed at different confidences.

2.2.2 Hypothesis 2: Systematic reviews produce a narrower range of opinion than those formed by participants only exposed to primary material

If exposure to a systematic review had an effect on opinion and more specifically that the opinions of participants become more consistent is tested. This prediction is based on the premise that if the participants find the evidential basis of a formalised review compelling then their opinions should migrate toward the findings of that review and thus become more consistent.

Again, the boot strapping approach is used in the analysis of this data. However, in answering this question instead of looking at the absolute differences between the graphs we compare several measures of statistical deviation (Standard deviation, Average deviation and Median average deviation). The distance measure being the sum of these deviation measures.

2.2.3 Hypothesis 3: Exposure to a professionally completed systematic review increases the perceived confidence participants have in their opinions

This hypothesis relates to the effect of a systematic review on the strength of opinion and tests whether participants are more confident in their opinions after viewing a systematic review. I.e. Is the population of confidences after exposure to a systematic review different to the population of confidences before exposure? The data collated to test this hypothesis is prepared and analysed using the same boot strapping techniques but applied solely to the expressed confidences. In this case the distance measure is the same three measures of statistical deviation but applied only to the expressed confidence.

Expressed certainty	Implied normalised standard deviation	Implied standard deviation over 9 point likert scale
1	2.997	26.97
2	1.347	12.12
3	0.605	5.45
4	0.272	2.45
5	0.122	1.10
6	0.055	0.49
7	0.025	0.22
8	0.011	0.10
9	0.005	0.04

Table 1 The standard deviations implied by expressed certainties

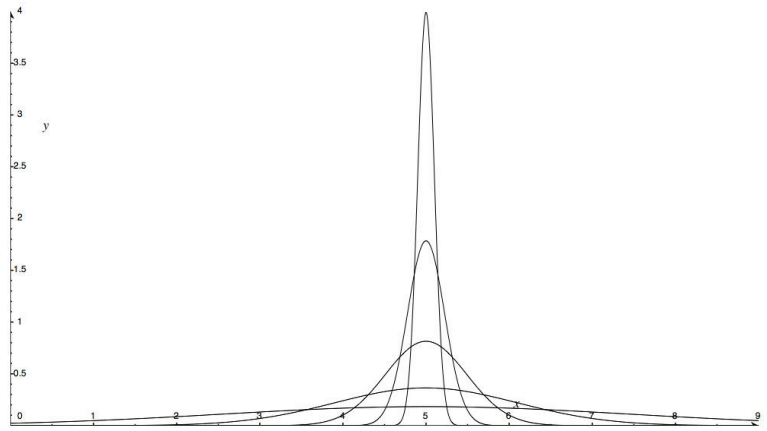

Figure 1 The probability density functions for an estimate of 5 based on expressed certainties of 4 for the flattest curve up to 8 for the steepest curve. Certainties of 1 to 3 are not shown as they are indistinguishable from 4. 9 is also omitted as it is functionally equivalent to 8.

3 RESULTS

The results section follows the methodology as detailed above. 51 subjects followed the protocol in 3 groups. Four Subjects declined to be included in the study, leaving 47 data sets to be analysed. In some instances, questions were not answered or marked as "not applicable" by the subjects. Where this has reduced the sample size for a calculation this has been noted in the results below.

3.1 Thematic analysis of open questions

Two of the questions on the surveys were open:

5) In what situation is pair programming particularly effective?
6) In what situation is pair programming particularly in-effective?

Independent experts were used, firstly to define classes of answers and then to assign the quantitative answers to these classes. The classes were derived independently of the question for which the answer was given. A further and mutually exclusive set of experts assigned scores on the nine-point Likert scale for each of these classes for both of the open questions. The classes and the number of answers in each class, mean and standard deviation for the Likert score assigned to each score are shown in table 2. We applied a standard deviation threshold of 1.5 amongst the opinion of the experts, answers falling into these categories were substituted by the Likert score from table 2. Other answers were not included in the analysis as there was no consensus as to their meaning amongst the expert panel.

	Question 5		Question 6	
	Mean	Std. Dev.	Mean	Std. Dev.
Middle Sized projects	6.46	2.67	2.02	3.92
No Cost Overhead	5.18	3.22	4.14	2.02
High end	8.01	1.02	3.09	0.80
Outline requirements	7.34	0.82	5.40	1.99
Rapid development	8.82	0.92	4.51	1.60
Small groups of experienced programmers	3.71	1.34	5.34	2.90
Inexperienced programmers	2.66	0.93	4.71	2.04
Low complexity systems	6.00	3.39	6.02	1.88
High complexity systems	7.10	1.62	7.83	2.89
Large systems	8.91	0.81	1.80	0.62
High requirement of success	6.02	2.02	3.90	1.04
Complex unknown team	2.00	2.99	2.07	3.99
Simple problem	1.04	.040	5.17	1.80
Most situations	9.00	0.00	1.00	0.00

Table 2 the results of the thematic analysis of the open questions

3.2 Questionnaire Data

Table 3 shows the average and standard deviation for the answers to questions 1 to 4 and for the scores derived from the open questions. Figure 6 shows the same data, but here it had been converted into probability distribution curves using the methodology discussed in section 2.2.1.

	After Study		Before Study	
	Mean	Std. Dev.	Mean	Std. Dev.
1	7.10	1.59	7.15	1.32
2	7.20	1.54	7.54	1.29
3	6.72	1.90	6.00	2.19
4	5.4.3218	1.56	5.18	2.82
5 & 6	6.63	2.11	5.12	2.19
Confidence	6.18	1.40	3.90	1.64

Table 3 mean and standard deviation for the answers to the questionnaire.

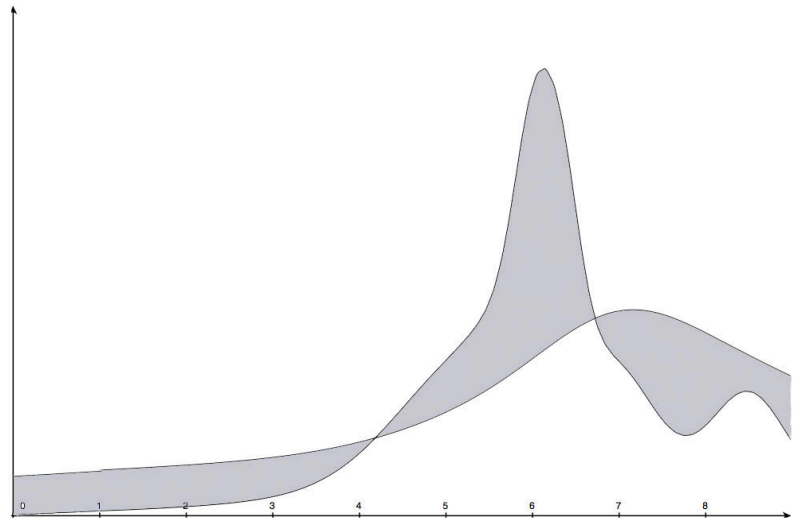

Figure 2 the sum of the probability distributions for the before and after data sets. The grey area highlights the difference or "distance" measure used in the statistical bootstrapping analysis. The x axis is each of the possible questionnaire answers in likert form. The y axis is the sum of the probability distribution curves for each participant.

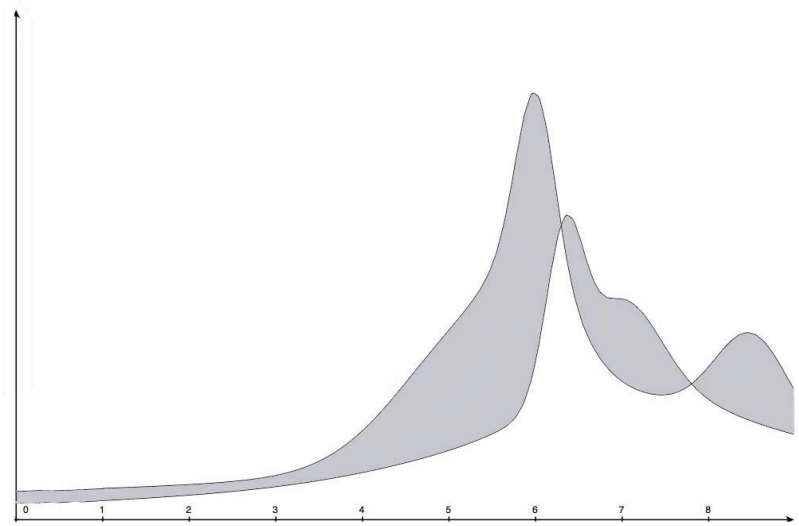

Figure 3 Probability distributions for one example of data sets created by random mixing as required by the statistical boot strapping analysis. The grey area highlights the difference or "distance". The x axis is each of the possible questionnaire answers in Likert form. The y axis is the sum of the probability distribution curves for each participant.

3.2.1 Hypothesis 1: Exposure to a systematic review will change the opinion of knowledgeable participants.

The data gathered when investigating this hypothesis falls into two populations, before and after. The null hypothesis is that these populations are subsets of the same population. I.e. that exposure to the study has not changed opinion.

The efficacy and confidence were calculated for each subject, both before and after the study. These values were used to produce a probability distribution curve representative of the 2 populations. Figure 6 shows the curves and highlights the difference or distance between them which is used in the bootstrapping analysis.

100,000 random mixings of the two populations were tested. In each case probability distribution curves representative of the two random populations were calculated and their distance measured. Figure 6 shows once such measure between two random mixings of the before and after populations. 3804 of these random mixings gave distance in excess of the one measured. Therefore, if the null hypothesis were true we would have expected to see an answer as extreme as the one we derived 3.804% of the time.

The null hypothesis associated with hypothesis 1 is shown to be false with a P value of 0.038, providing good evidence for accepting hypothesis 1 as true.

3.2.2 Hypothesis 2: Systematic reviews produce a narrower range of opinion than those formed by participants only exposed to primary material.

The calculated efficacies of the before and after populations were used to test the null hypothesis, that "Exposure to the study does not cause a narrowing of opinion". The fall in standard deviation, average deviation and median average deviation was measured between the two populations. 100,000 random mixings of the two populations where then measured and any fall in standard deviation, average deviation and median average deviation recorded. Of the 300,000 measurements 82952 recoded a fall greater than the ones between the original populations. Therefore, if the null hypothesis were true we would have expected to see an answer as extreme at the one we derived 27.65% of the time.

The experiment provides little evidence for the null hypothesis and no indication for the validity of hypothesis 2.

3.2.3 Hypothesis 3: Exposure to a professionally completed systematic review increases the perceived confidence

The confidence of the participants in the before and after populations were used to evaluate the null hypothesis, that "Exposure to the study does not cause confidence to rise". The rise in mean, median and mode of the expressed confidence was measured between the two populations. 100,000 random mixings of the two populations where then measured and any raise in mean, median and mode recorded. Of the 300,000 measurements 2907 recoded a rise greater than the ones between the original populations. Therefore, if the null hypothesis were true we would have expected to see an answer as extreme as the one we derived 0.96% of the time.

The null hypothesis associated with hypothesis 3 is shown to be false with a P value of 0.01, providing excellent evidence for the assumption that hypothesis 3 is true.

4 Discussion

This paper reports an experiment that investigated the effects of exposing practicing engineers to evidence-based literature reviews. We want to understand if these reviews have the power to effect practice without their findings being mandated. Work in the medical sciences has shown that there is a small, but positive effect (Antman 1992) and researchers in EBSE have argued that there should be a similar change for practising software engineers. It could even be argued that the effects should be stronger in this field. EBSE is a method for constructing an argument based on evidence derived from primary studies. Computer scientists and software engineers have training that is highly logical and the discipline attracts people with a logical thought process. Shouldn't the logic of EBSE be more attractive and therefore have more effect on the discipline's practitioners?

Our goal is to understand if practice is changed. To address this question directly would take a longitudinal study where we investigate the outcomes of projects and the long-term adoption or rejection of the findings of studies. Here we meet probably the most complex issue faced by EBSE, the lack of repeatable projects to trial our ideas. Unlike medicine, the software industry has a relatively small number of large projects, which makes the statistical analysis and double-blind trial methodology impossible. Without the ability to directly test our objective, we approach the next best question; does EBSE effect opinion? The assumption that we are making is that if opinion is effected then practice will be effected when these engineers return to their projects. One future line of enquiry will be to look for differences in adoption of technologies between groups that received the EBSE training and those that did not. However, we hold a low expectation that sufficient subjects will be available to produce a statistically significant result.

The discussion is split into two sections, a discussion of the result and one on the methodology. We believe that both are novel and warrant a detailed analysis.

4.1 Discussion on results

Our experiments provide excellent evidence that systematic reviews have the power to change opinion, even amongst mid career engineers who have already been fully briefed on the technologies being discussed in the review. Anecdotally, the subjects referred to the systematic review as neutral rather than as partisan. It was this neutrality and the logical laying out of the evidence in both directions that was appealing, rather than the statistical analysis performed by the EBSE authors.

While a positive result, the feedback brings into question the value of the EBSE study. It appears to have been received more as a laying out of the evidence than as an objective analysis. However, there is little doubt that the opinions were effected by the EBSE study, which validates the efforts of the software engineering community to produce these studies.

The experiment also showed that exposure to the EBSE study made a large impact on the confidence of the participants in their own understanding. We believe that this is a major outcome of both the

EBSE studies and this experiment. While very few people's opinion moved hugely, many people become confident in their understanding to the point where their ability and willingness to adopt the technologies reached a threshold level. Again, in feedback after the last phase of the project, several participants commented that they understood the opinion of the research community before the study, that their initial training had delivered the facts. But, that the EBSE work had reinforced their conclusions and bolstered their confidence to the point where they now planned to attempt to use it rather than consider the technology as another "academic idea with little real life relevance"

We also expected the study to show that opinion narrowed as a result of the study. I.e. that everybody would move closer to the conclusions of the study. However, we were surprised that this did not appear to be the case. The graphs showed that prior to the study there was a uni-modal distribution to the opinion, largely driven by the low levels of confidence. After the study opinion had hardened and became bi-modal in distribution. On reflection there are several reasons that could explain this. Firstly, pair programming is a hotly debated technique and all opinion is divided. Even the systematic review does not give a definitive answer. Secondly, we believe that the subjects fell into two categories, people who see pair programming as a useful tool but not relevant to their projects and those that see it as directly relevant to their projects. Either or both of these reasons could account for the lack of a narrowing of opinion.

4.2 Discussion on method

The methods we have adopted in analysing our findings bears some discussion. The use of Likert scales in questionnaires is a common technique when measuring opinion and attitudes (Dawes 2008). Neuman argues that its popularity may be due to the ease in which these scales are developed (Neuman 1994) and a body of research which has demonstrated that these scales tend to be more reliable than some other scales with the same number of items. However, in the present study we were interested in the confidence that participants had when expressing their answers in terms of how certain they were of the answers expressed. To our knowledge there is no work available within the field of psychology that provides a model of how confidence in people's answers to questions can be represented. That is, we could find no model that integrated a representation of certainty into a test that could be applied to a hypothesis. As a result, we have developed our own model of confidence and built the analysis of our findings around it.

The finding of most interest here is that on a nine-point scale the participants were really working on at most five points. Responses given in the range of one to four were functionally equivalent, in that there was little discrimination between the possible values and this was also the case for responses of eight and nine on a nine-point scale. Therefore, it is not clear what the difference is between a participant giving 8/9 or 9/9 as an expressed level of confidence. Though this was only when expressing confidence, this could call into question research whose experimental basis is the analysis of data captured using Likert scales. These numerical scales represent one type of response scale. Also common are scales that use only verbal anchors. The findings suggest that the adoption of a Likert scale with a smaller range of numerical responses i.e. a five-point or seven-point scale may be more reliable in measuring efficacy, where each value along the scale is mapped onto a single verbal response. For example, a scale of one to five may reflect Strongly agree (1), agree (2), neither agree nor disagree (3), disagree (4) and strongly disagree (5).

5 Conclusions

Kitchenham et. al. stated in a very early paper introducing EBSE that systematic review should be quickly adopted within software engineering (Kitchenham 2004) . Jrgensen et. al. agreed (Jorgensen 2005) but admit that while evidence based science has been shown to improve outcomes in medicine, there is no direct evidence for its effects in software engineering. This paper has shown that EBSE does have a real and measurable effect on practicing engineers.

Our conclusions can be summarised as follows:

1. Showing a systematic review to an experienced engineer can change their opinion, even if they have direct experience and training on the subject of the review.

2. Reviews mostly change opinion by improving confidence in the subjects rather by making large changes to their base opinion. This then supports the engineer in adopting or rejecting the technology in an active manner.

6 References

Antman, E.M., Lau,J., Kupelnick,B., Mosteller,F. and Chalmers,T.C. (1992) A comparison of results of meta-analysis of randomized controlled trials and recommendations of clinical experts. JAMA-Journal of the American Medical Association, 268:240–248.

Dawes,J. (2008) Do data characteristics change according to the number of scale points used? an experiment using 5-point, 7-point and 10-point scales. International journal of Marke Research, 50(1).

Hannay,J.E., Dyb,T., Arisholm,E., and Sjberg,D.I.K.(2009). The effectiveness of pair programming: A meta-analysis. Information and Software Technology, 51(7):1110 – 1122. Special Section: Software Engineering for Secure Systems.

Jorgensen,M., Dybo,T., and Kitchenham,B. (2005) Teaching evidence-based software engineering to university students. Software Metrics, IEEE International Symposium on, 0:24.

Kitchenham, B., Brereton,O.P., Budgen, D., Turner, M., Bailey,J., and Linkman,S. (2009) Systematic literature reviews in software engineering - a systematic literature review. Information and Software Technology, 51(1):7 – 15

.

Kitchenham, B., Dyba, T. and Jorgensen,M (2004) Evidence-based software engineering. In ICSE '04: Proceedings of the 26th International Conference on Software Engineering, pages 273–281, Washington, DC, USA. IEEE Computer Society.

Neuman, W.L. (1994) Social Research Methods: Qualitive and Quantitative Approaches. USA:Allyn & Bacon.